Asymptotic properties of nonlinear estimates in stochastic models with finite design space[☆]

Luc Pronzato

Laboratoire I3S, CNRS/Université de Nice-Sophia Antipolis
Bât. Euclide, Les Algorithmes, 2000 route des lucioles, BP 121
06903 Sophia Antipolis cedex, France

Abstract

Under the condition that the design space is finite, new sufficient conditions for the strong consistency and asymptotic normality of the least-squares estimator in nonlinear stochastic regression models are derived. Similar conditions are obtained for the maximum-likelihood estimator in Bernoulli type experiments. Consequences on the sequential design of experiments are pointed out.

Key words: stochastic regressors, strong consistency, asymptotic normality, sequential design, Bernoulli trials

1. Introduction and motivation

Consider a nonlinear regression model with observations

$$Y_i = Y(x_i) = \eta(x_i, \bar{\theta}) + \varepsilon_i \,, \tag{1}$$

where $\{\varepsilon_i\}$ is a martingale difference sequence with respect to an increasing sequence of σ-fields $\mathcal{F}_i$ such that $\sup_i \mathbb{E}\{\varepsilon_i^2 | \mathcal{F}_{i-1}\} < \infty$ almost surely (a.s.), and $\eta(x, \theta)$ is a known function of a parameter vector $\theta \in \Theta$ (a compact subset of $\mathbb{R}^p$) and a design variable $x \in \mathcal{X}$ (a compact subset of $\mathbb{R}^d$). Here $\bar{\theta}$ denotes the unknown true value of θ and we assume that $\bar{\theta}$ is in the interior of Θ. The martingale difference sequence assumption

[☆]This work was partly accomplished while the author was invited at the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK, in July 2008. The support of the Newton Institute and of CNRS are gratefully acknowledged.

Email address: pronzato@i3s.unice.fr (Luc Pronzato)

for $\{\varepsilon_i\}$ in (1) is rather common in a stochastic control framework. It covers situations where $\varepsilon_i = h_i\,\delta_i$ with h_i being a measurable function of past ε and $\{\delta_i\}$ forming an i.i.d. sequence with zero mean also independent of past ε. A typical example is given by ARCH (autoregressive conditionally heteroscedastic) processes.

The strong consistency of the Least-Squares (LS) estimator $\hat{\theta}^n$ that minimizes

$$S_n(\theta) = \sum_{k=1}^{n} [Y(x_k) - \eta(x_k, \theta)]^2 \tag{2}$$

is established in (Jennrich, 1969) in the case where ε_i are independent identically distributed (i.i.d.) errors with unknown variance σ^2 *and* x_i *are non-random constants*, under the assumption that $(1/n)D_n(\theta, \theta')$ converges uniformly to a continuous function $J(\theta, \theta')$ with $J(\theta, \theta') > 0$ for all $\theta \neq \theta'$, where

$$D_n(\theta, \theta') = \sum_{i=1}^{n} [\eta(x_i, \theta) - \eta(x_i, \theta')]^2 \,. \tag{3}$$

In a linear regression model, where $\eta(x, \theta) = \mathbf{f}^\top(x)\theta$ with $\mathbf{f}(x)$ a p-dimensional vector, the condition above is equivalent to $(1/n)\mathbf{X}_n^\top \mathbf{X}_n \to \mathbf{M}$, with $\mathbf{M}$ some positive definite matrix and $\mathbf{X} = [\mathbf{f}(x_1), \ldots, \mathbf{f}(x_n)]^\top$, a condition thus much stronger than the well-known condition for weak and strong consistency of $\hat{\theta}^n$

$$(\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \to 0 \,, \tag{4}$$

see, e.g., Lai et al. (1978); Lai and Wei (1982). The analogue of (4) for nonlinear regression would be $D_n(\theta, \theta') \to \infty$ for all $\theta \neq \theta'$. This condition is shown in (Wu, 1981) to be necessary for the existence of a weakly consistent estimator of θ when ε_i are supposed to be i.i.d. with a positive almost everywhere and absolutely continuous density with finite Fisher information. It is also shown in the same paper to be sufficient for the (weak and strong) consistency of the nonlinear LS estimator $\hat{\theta}^n$ when Θ is a finite set. When Θ is a compact set of $\mathbb{R}^p$, it is complemented by additional assumptions to establish the strong consistency of $\hat{\theta}^n$, see Th. 3 in (Wu, 1981).

Suppose now that x_i *is a* $\mathcal{F}_{i-1}$ *measurable random variable.* The motivation we have in mind corresponds to adaptive experimental design, where the design point x_i at step i

depends on observations $Y_1, \dots, Y_{i-1}$ through the estimate $\hat{\theta}^{i-1}$. Lai and Wei (1982) show that the conditions

$$\lambda_{\min}[\mathbf{X}_n^\top \mathbf{X}_n] \to \infty \quad \text{a.s.} \tag{5}$$

$$\{\log \lambda_{\max}[\mathbf{X}_n^\top \mathbf{X}_n]\}^\rho = o\left(\lambda_{\min}[\mathbf{X}_n^\top \mathbf{X}_n]\right) \quad \text{a.s.} \quad \text{for some } \rho > 1\,, \tag{6}$$

are sufficient for the strong consistency of $\hat{\theta}^n$ in the model (1) with $\eta(x, \theta)$ linear in θ, i.e. $\eta(x, \theta) = \mathbf{f}^\top(x)\theta$, and stochastic regressors $\mathbf{f}(x_i)$ (Example 1 in the same paper shows that these conditions are in some sense weakest possible). Here and in what follows we denote by $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ the minimum and maximum eigenvalues of a $p \times p$ matrix $\mathbf{M}$. The case of nonlinear stochastic regression models is considered in (Lai, 1994), where sufficient conditions for strong consistency are given, which reduce to (5) and the Christopeit and Helmes (1980) condition, $\lambda_{\max}[\mathbf{X}_n^\top \mathbf{X}_n] = \mathcal{O}\{\lambda_{\min}^\rho[\mathbf{X}_n^\top \mathbf{X}_n]\}$ a.s. for some $\rho \in (1, 2)$, in the case of a linear model.

It is the purpose of this paper to show that *when the design space $\mathcal{X}$ is finite*, a sufficient condition for the strong consistency of $\hat{\theta}^n$ in the model (1) is that with probability one $D_n(\theta, \theta') \to \infty$ faster than $(\log n)^\rho$ for all $\theta \neq \theta'$ for some $\rho > 1$, a condition equivalent to (5, 6) for linear models and much weaker than the conditions of Jennrich (1969) or Lai (1994) for nonlinear models. Under the additional assumption

$$\lim_{i \to \infty} \mathbb{E}\{\varepsilon_i^2 | \mathcal{F}_{i-1}\} = \sigma^2 \quad \text{a.s.} \quad \text{for some constant } \sigma\,, \tag{7}$$

we also give a sufficient condition for the asymptotic normality of $\hat{\theta}^n$ in (1). It should be noticed that the assumption that $\mathcal{X}$ is finite is seldom limitative in situations where the experiment is designed since practical considerations often impose such a restriction on possible choices for x_i. This is especially true for clinical trials where only certain doses of the treatment are available, see Sect. 4 and Pronzato (2009b). Although less natural in a stochastic control context where x_i denotes the system input at time i, the assumption that $\mathcal{X}$ is finite is satisfied when a suitable quantization is applied to the input sequence. It can be contrasted with the less natural assumption that the admissible parameter set Θ is finite, see, e.g., Caines (1975).

Sect. 2 concerns the strong consistency of $\hat{\theta}^n$ and Sect. 3 its asymptotic normality. The results obtained, which rely on a repeated sampling principle that can be used when $\mathcal{X}$ is finite, are of rather general applicability and Sect. 4 concerns Maximum Likelihood (ML) estimation in Bernoulli trials. Sect. 5 concludes and points out consequences on sequentially designed experiments. We respectively denote $\overset{\text{a.s.}}{\to}$, $\overset{\text{p}}{\to}$ and $\overset{\text{d}}{\to}$ almost sure convergence, convergence in probability and in distribution. For $\mathbf{M}$ a $p \times p$ matrix, we use the matrix norm $\|\mathbf{M}\| = \sup_{\|\mathbf{u}\|=1} \|\mathbf{M}\mathbf{u}\| \leq p \max_{i,j} |\{\mathbf{M}\}_{ij}|$.

2. Strong consistency of the nonlinear LS estimator when $\mathcal{X}$ is finite

Next theorem shows that the strong consistency of $\hat{\theta}^n$ in (1) is a consequence of $D_n(\theta, \bar{\theta})$ tending to infinity fast enough for $\|\theta - \bar{\theta}\| \geq \delta > 0$. The fact that the design space $\mathcal{X}$ is finite makes the required rate of increase for $D_n(\theta, \bar{\theta})$ quite slow. The result is valid whether x_i are non-random constants or are $\mathcal{F}_{i-1}$-measurable random variables.

Theorem 1. *Suppose that $\mathcal{X}$ is a finite set. If $D_n(\theta, \bar{\theta})$ given by (3) satisfies*

$$\text{for all } \delta > 0\,, \ \left[\inf_{\|\theta-\bar{\theta}\|\geq\delta/\tau_n} D_n(\theta,\bar{\theta})\right] / (\log n)^\rho \overset{\text{a.s.}}{\to} \infty \ (n \to \infty)\,, \quad \text{for some } \rho > 1\,, \tag{8}$$

with $\{\tau_n\}$ a nondecreasing sequence of positive deterministic constants, then the LS estimator $\hat{\theta}^n$ in the model (1) satisfies

$$\tau_n \|\hat{\theta}^n - \bar{\theta}\| \overset{\text{a.s.}}{\to} 0 \ (n \to \infty)\,. \tag{9}$$

Proof. The first part of the proof is based on Lemma 1 in (Wu, 1981). Suppose that (9) is not satisfied. It implies that exists $\delta > 0$ such that

$$\Pr(\limsup_{n\to\infty} \tau_n \|\hat{\theta}^n - \bar{\theta}\| \geq \delta) > 0\,. \tag{10}$$

Since $S_n(\hat{\theta}^n) \leq S_n(\bar{\theta})$, (10) implies $\Pr(\liminf_{n\to\infty} \inf_{\|\theta-\bar{\theta}\|\geq\delta/\tau_n} [S_n(\theta) - S_n(\bar{\theta})] \leq 0) > 0$. Therefore,

$$\liminf_{n\to\infty} \inf_{\|\theta-\bar{\theta}\|\geq\delta/\tau_n} [S_n(\theta) - S_n(\bar{\theta})] > 0 \ \text{ a.s. for any } \delta > 0 \tag{11}$$

implies (9). The second part consists in establishing a sufficient condition for (11) based on the growth rate of $D_n(\theta,\bar{\theta})$. Denote $\mathcal{I}_n(x) = \{i \in \{1,\dots,n\} : x_i = x\}$. We have

$$S_n(\theta) - S_n(\bar{\theta}) \;\geq\; D_n(\theta,\bar{\theta}) \left[1 - 2\,\frac{\sum_{x\in\mathcal{X}} \left|\sum_{i\in\mathcal{I}_n(x)} \varepsilon_i\right| |\eta(x,\bar{\theta}) - \eta(x,\theta)|}{D_n(\theta,\bar{\theta})}\right].$$

Under the condition (8), it thus suffices to prove that

$$\limsup_{n\to\infty} \sup_{\|\theta-\bar{\theta}\|\geq\delta/\tau_n} \frac{\sum_{x\in\mathcal{X}} \left|\sum_{i\in\mathcal{I}_n(x)} \varepsilon_i\right| |\eta(x,\bar{\theta}) - \eta(x,\theta)|}{D_n(\theta,\bar{\theta})} = 0 \;\text{ a.s. for any } \delta > 0 \qquad (12)$$

to obtain (11) and thus (9). Denote $u_i(x)$ the variable defined by $u_i(x) = 1$ if $x = x_i$ and $u_i(x) = 0$ otherwise, so that $\sum_{i=1}^n u_i(x) = \sum_{i=1}^n u_i^2(x) = r_n(x)$, the number of times x appears in the sequence $x_1,\dots,x_n$. Notice that $u_i(x)$ is $\mathcal{F}_{i-1}$-measurable. Since $D_n(\theta,\bar{\theta}) \geq D_n^{1/2}(\theta,\bar{\theta}) r_n^{1/2}(x) |\eta(x,\bar{\theta}) - \eta(x,\theta)|$ for all x in $\mathcal{X}$, we have

$$\frac{\sum_{x\in\mathcal{X}} \left|\sum_{i\in\mathcal{I}_n(x)} \varepsilon_i\right| |\eta(x,\bar{\theta}) - \eta(x,\theta)|}{D_n(\theta,\bar{\theta})} \leq \frac{1}{D_n^{1/2}(\theta,\bar{\theta})} \sum_{x\in\mathcal{X}} \frac{|\sum_{i=1}^n u_i(x)\varepsilon_i|}{[\sum_{i=1}^n u_i^2(x)]^{1/2}}.$$

Moreover, $A_n(x) = |\sum_{i=1}^n u_i(x)\varepsilon_i| \,[\sum_{i=1}^n u_i^2(x)]^{-1/2}$ is a.s. finite if $r_n(x)$ is finite and

$$\lim_{n\to\infty} \frac{|\sum_{i=1}^n u_i(x)\varepsilon_i|}{[\sum_{i=1}^n u_i^2(x)]^{1/2} [\log \sum_{i=1}^n u_i^2(x)]^{\alpha}} = 0 \;\text{ a.s.}$$

for every $\alpha > 1/2$ otherwise, see Lemma 2-(iii) of Lai and Wei (1982) and Corollary 7 of Chow (1965). Since $\sum_{i=1}^n u_i^2(x) \leq n$ for all x, (8) implies (12), which concludes the proof.

■

Remark 1. The case where the errors ε_i in (1) are i.i.d. with finite variance σ^2 is considered in (Pronzato, 2009a). $A_n(x)$ is then asymptotically normal $\mathcal{N}(0,\sigma^2)$ when $r_n(x) \to \infty$ and, using the law of the iterated logarithm, we obtain (9) under the weaker condition

$$\text{for all } \;\delta > 0\,,\; \left[\inf_{\|\theta-\bar{\theta}\|\geq\delta/\tau_n} D_n(\theta,\bar{\theta})\right] / (\log\log n) \overset{\text{a.s.}}{\to} \infty \;(n\to\infty)\,, \qquad (13)$$

see also Th. 3 below. Under the same assumption of i.i.d. errors with finite variance and using a similar approach, we also obtain in (Pronzato, 2009a) that $\hat{\theta}^n$ is weakly consistent

when $D_n(\theta,\bar\theta) \xrightarrow{\mathrm{p}} \infty$ for all $\theta \neq \bar\theta$ as $n \to \infty$, which can be extended to $\tau_n\|\hat\theta^n - \bar\theta\| \xrightarrow{\mathrm{p}} 0$ when $\inf_{\|\theta-\bar\theta\|\geq\delta/\tau_n} D_n(\theta,\bar\theta) \xrightarrow{\mathrm{p}} \infty$ for all $\delta > 0$. The same property still holds when $\{\varepsilon_i\}$ in (1) is a martingale difference sequence that satisfies (7) and $r_n(x)$, the number of times x appears in the sequence $x_1,\ldots,x_n$, satisfies $r_n(x)/\mathbb{E}\{r_n(x)\} \xrightarrow{\mathrm{p}} 1$ for all $x \in \mathcal{X}$. In that case, (9) can be obtained under a slightly weaker condition than (8) using results on the law of the iterated logarithm for martingales, see Hall and Heyde (1980, Chap. 4).

3. Asymptotic normality of the nonlinear LS estimator when $\mathcal{X}$ is finite

We make the following regularity assumption on the model response $\eta(x,\theta)$ in (1):

$\mathbf{H}_\eta$: $\eta(x,\theta)$ is two times continuously differentiable with respect to θ in some open neighborhood of $\bar\theta$ for all $x \in \mathcal{X}$.

We denote $\mathbf{f}_\theta(x) = \partial\eta(x,\theta)/\partial\theta$ and

$$\mathbf{M}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{f}_\theta(x_i)\mathbf{f}_\theta^\top(x_i)\,. \tag{14}$$

Theorem 2. *Suppose that $\mathcal{X}$ is a finite set, that the errors ε_i in (1) satisfy (7) and that $\eta(x,\theta)$ satisfies the regularity condition $\mathbf{H}_\eta$. Suppose that there exist non-random symmetric positive definite $p \times p$ matrices $\mathbf{C}_n$ such that*

$$\mathbf{C}_n^{-1}\mathbf{M}_n^{1/2}(\bar\theta) \xrightarrow{\mathrm{p}} \mathbf{I}\,, \tag{15}$$

with $\mathbf{I}$ the p-dimensional identity matrix, and that $c_n = \lambda_{\min}(\mathbf{C}_n)$ and $D_n(\theta,\bar\theta)$ satisfy

$$n^{1/4}c_n \to \infty \text{ and } \forall\delta > 0\,,\ \inf_{\|\theta-\bar\theta\|\geq c_n^2\delta} D_n(\theta,\bar\theta)/(\log n)^\rho \xrightarrow{\text{a.s.}} \infty \text{ for some } \rho > 1\ (n \to \infty)\,. \tag{16}$$

Then the LS estimator $\hat\theta^n$ in the model (1) satisfies

$$\sqrt{n}\,\mathbf{M}_n^{1/2}(\hat\theta^n)(\hat\theta^n - \bar\theta) \xrightarrow{\mathrm{d}} \omega \sim \mathcal{N}(0,\sigma^2\mathbf{I})\,,\ n \to \infty\,. \tag{17}$$

Proof. Since $\mathcal{X}$ is finite, c_n is bounded from above and (16) implies $\hat\theta^n \xrightarrow{\text{a.s.}} \bar\theta$, see Th. 1. Therefore, there exists a ball $\mathcal{B}(\bar\theta,r)$ centered at $\bar\theta$, included in Θ and such that $\hat\theta^n \in \mathcal{B}(\bar\theta,r)$

for all n larger than some N_0. We can thus consider a first-order series expansion of $\partial S_n(\theta)/\partial\theta$ around $\bar{\theta}$, with $S_n(\theta)$ given by (2). This yields

$$\frac{\partial S_n(\theta)}{\partial \theta_j}\Big|_{\hat{\theta}^n} = 0 = \frac{\partial S_n(\theta)}{\partial \theta_j}\Big|_{\bar{\theta}} + (\hat{\theta}^n - \bar{\theta})^\top \frac{\partial^2 S_n(\theta)}{\partial\theta\partial\theta_j}\Big|_{\tilde{\theta}_j^n}, \; j = 1, \ldots, p\,,$$

where $\tilde{\theta}_j^n$ denotes some value between $\hat{\theta}_j^n$ and $\bar{\theta}_j$. Direct calculations give $\sum_{i=1}^n \varepsilon_i \mathbf{f}_{\bar{\theta}}(x_i) = n\mathbf{M}_n(\bar{\theta})(\hat{\theta}^n - \bar{\theta}) + n(\mathbf{R}_{n,1} + \mathbf{R}_{n,2} + \mathbf{R}_{n,3})(\hat{\theta}^n - \bar{\theta})$ with $\{\mathbf{R}_{n,1}\}_{j,k} = \{\mathbf{M}_n(\tilde{\theta}_j^n) - \mathbf{M}_n(\bar{\theta})\}_{jk}$, $\{\mathbf{R}_{n,2}\}_{j,k} = -(1/n)\sum_{i=1}^n \varepsilon_i\, \partial^2\eta(x_i,\theta)/(\partial\theta_j\partial\theta_k)|_{\tilde{\theta}_j^n}$, $\{\mathbf{R}_{n,3}\}_{j,k} = (1/n)\sum_{i=1}^n [\eta(x_i,\tilde{\theta}_j^n) - \eta(x_i,\bar{\theta})]\, \partial^2\eta(x_i,\theta)/(\partial\theta_j\partial\theta_k)|_{\tilde{\theta}_j^n}$. We thus obtain

$$\frac{1}{\sqrt{n}}\mathbf{C}_n^{-1}\sum_{i=1}^n \varepsilon_i \mathbf{f}_{\bar{\theta}}(x_i) = \mathbf{C}_n^{-1}\left[\mathbf{M}_n(\bar{\theta}) + \mathbf{R}_{n,1} + \mathbf{R}_{n,2} + \mathbf{R}_{n,3}\right]\mathbf{C}_n^{-1}\,\mathbf{C}_n\sqrt{n}(\hat{\theta}^n - \bar{\theta})\,, \qquad (18)$$

where $\mathbf{C}_n^{-1}\mathbf{M}_n(\bar{\theta})\mathbf{C}_n^{-1} \xrightarrow{\mathrm{p}} \mathbf{I}$ from (15). Consider the three terms $\mathbf{C}_n^{-1}\mathbf{R}_{n,j}\mathbf{C}_n^{-1}$, $j = 1, 2, 3$. We have $\|\mathbf{C}_n^{-1}[\mathbf{M}_n(\theta) - \mathbf{M}_n(\bar{\theta})]\mathbf{C}_n^{-1}\| \leq (p/c_n^2)\,\max_{j,k}|\{\mathbf{M}_n(\theta) - \mathbf{M}_n(\bar{\theta})\}_{jk}|$ and thus

$$\|\mathbf{C}_n^{-1}[\mathbf{M}_n(\theta) - \mathbf{M}_n(\bar{\theta})]\mathbf{C}_n^{-1}\| \leq \frac{p}{c_n^2}\max_{j,k}\max_{x\in\mathcal{X}}|\{\mathbf{f}_\theta(x)\}_j\{\mathbf{f}_\theta(x)\}_k - \{\mathbf{f}_{\bar{\theta}}(x)\}_j\{\mathbf{f}_{\bar{\theta}}(x)\}_k| \leq \frac{A}{c_n^2}\,\|\theta - \bar{\theta}\|$$

for some $A > 0$. Therefore, $\|\mathbf{C}_n^{-1}\mathbf{R}_{n,1}\mathbf{C}_n^{-1}\| \xrightarrow{\mathrm{p}} 0$ as $n \to \infty$ (using (16) and Th. 1). For the second term we obtain

$$\left\|\mathbf{C}_n^{-1}\frac{1}{n}\sum_{i=1}^n \varepsilon_i \frac{\partial^2\eta(x_i,\theta)}{\partial\theta\partial\theta^\top}\,\mathbf{C}_n^{-1}\right\| \;\leq\; \frac{1}{c_n^2\sqrt{n}}\sum_{x\in\mathcal{X}}\frac{|\sum_{i=1}^n u_i(x)\varepsilon_i|}{\sqrt{n}}\max_{x\in\mathcal{X},\theta\in\Theta}\left\|\frac{\partial^2\eta(x_i,\theta)}{\partial\theta\partial\theta^\top}\right\|$$

where $u_i(x) = 1$ if $x = x_i$ and $u_i(x) = 0$ otherwise. Since $\mathbb{E}(\varepsilon_i^2|\mathcal{F}_{i-1}) < \infty$ a.s., we obtain that $|\sum_{i=1}^n u_i(x)\varepsilon_i|/\sqrt{n}$ is bounded in probability for all x. Therefore, $n^{1/4}c_n \to \infty$ implies $\|\mathbf{C}_n^{-1}\mathbf{R}_{n,2}\mathbf{C}_n^{-1}\| \xrightarrow{\mathrm{p}} 0$ as $n \to \infty$. Finally, we get for the third term,

$$\left\|\mathbf{C}_n^{-1}\frac{1}{n}\sum_{i=1}^n[\eta(x_i,\theta) - \eta(x_i,\bar{\theta})]\,\frac{\partial^2\eta(x_i,\theta)}{\partial\theta\partial\theta^\top}\,\mathbf{C}_n^{-1}\right\|$$
$$\leq \frac{p}{c_n^2}\max_{j,k}\max_{x\in\mathcal{X}}\left[|\eta(x,\theta) - \eta(x,\bar{\theta})|\max_{\theta\in\Theta}\left|\frac{\partial^2\eta(x,\theta)}{\partial\theta_j\partial\theta_k}\right|\right] \leq \frac{B}{c_n^2}\,\|\theta - \bar{\theta}\|$$

for some $B > 0$, and $\|\mathbf{C}_n^{-1}\mathbf{R}_{n,3}\mathbf{C}_n^{-1}\| \xrightarrow{\mathrm{p}} 0$ as $n \to \infty$ for the same reasons as for $\mathbf{R}_{n,1}$.

Substitution in (18) yields $(1/\sqrt{n})\,\mathbf{C}_n^{-1}\sum_{i=1}^n \varepsilon_i\,\mathbf{f}_{\bar{\theta}}(x_i) = [1 + o_p(1)]\mathbf{C}_n\sqrt{n}(\hat{\theta}^n - \bar{\theta})$ and thus, using (15), $\sqrt{n}\mathbf{M}_n^{1/2}(\hat{\theta}^n)(\hat{\theta}^n - \bar{\theta}) = [1 + o_p(1)](1/\sqrt{n})\,\mathbf{C}_n^{-1}\sum_{i=1}^n \varepsilon_i\,\mathbf{f}_{\bar{\theta}}(x_i)$. Since we

have $\max_i(1/\sqrt{n})\,\|\mathbf{C}_n^{-1}\mathbf{f}_{\bar\theta}(x_i)\| \leq [1/(c_n\sqrt{n})]\,\max_x\|\mathbf{f}_{\bar\theta}(x)\| \to 0$, $\mathbb{E}(\varepsilon_i^2|\mathcal{F}_{i-1}) \overset{\text{a.s.}}{\to} \sigma^2$ and $\mathbf{C}_n^{-1}(1/n)\left[\sum_{i=1}^n \mathbf{f}_{\bar\theta}(x_i)\mathbf{f}_{\bar\theta}^\top(x_i)\right]\mathbf{C}_n^{-1} \overset{\text{p}}{\to} \mathbf{I}$, we are in the same situation as in (Lai, 1994, Th.2) and (17) follows from the martingale central limit Theorem. Indeed, consider $T_n = (1/\sqrt{n})\mathbf{u}^\top\mathbf{C}_n^{-1}\sum_{i=1}^n \varepsilon_i\mathbf{f}_{\bar\theta}(x_i)$, with $\mathbf{u}$ any vector of $\mathbb{R}^p$ with norm 1. The conditional Lindeberg condition and the condition on conditional variances in (Dvoretzky, 1972, Th. 2.2) are satisfied and T_n is asymptotically normal $\mathcal{N}(0,\sigma^2)$. ∎

Remark 2.

(i) One may notice that compared to (Wu, 1981), we do not require that $(n/\tau_n)\mathbf{M}_n(\bar\theta)$ tends to some positive definite matrix for some $\tau_n \to \infty$ and, compared to (Lai, 1994) we do not require the existence of high-order derivatives of $\eta(x,\theta)$. On the other hand, we suppose that $\mathcal{X}$ is finite and we need that $c_n = \lambda_{\min}(\mathbf{C}_n)$ decreases more slowly than $n^{-1/4}$, see (16) (one may notice that when $\mathcal{X}$ is finite, the condition (2.5) of Lai (1994) imposes that c_n is bounded from below).

(ii) When ε_i in (1) are i.i.d. with finite variance σ^2, the condition (16) can be replaced by $n^{1/4}c_n \to \infty$, $\hat\theta^n \overset{\text{a.s.}}{\to} \bar\theta$ and $\inf_{\|\theta-\bar\theta\|\geq c_n^2\delta} D_n(\theta,\bar\theta) \overset{\text{p}}{\to} \infty$ for all $\delta > 0$, see Remark 1. Indeed, this is enough to obtain $\|\hat\theta^n - \bar\theta\|/c_n^2 \overset{\text{p}}{\to} 0$, which implies that $\|\mathbf{C}_n^{-1}\mathbf{R}_{n,1}\mathbf{C}_n^{-1}\| \overset{\text{p}}{\to} 0$ and $\|\mathbf{C}_n^{-1}\mathbf{R}_{n,3}\mathbf{C}_n^{-1}\| \overset{\text{p}}{\to} 0$ as $n \to \infty$.

(iii) When x_i is $\mathcal{F}_{i-1}$ measurable, $\mathbf{M}_n(\theta)$ is not in general the information matrix for parameters θ. This is true in particular for sequential experimental design. Under the assumptions of the theorem, it is legitimate, however, to characterize the asymptotic precision of the estimation by $\mathbf{M}_n^{-1}(\hat\theta^n)$. For instance, in the case of sequential D-optimal design where $x_{n+1} = \arg\max_{x\in\mathcal{X}} \mathbf{f}_{\hat\theta^n}^\top(x)\mathbf{M}_n^{-1}(\hat\theta^n)\mathbf{f}_{\hat\theta^n}(x)$, it is shown in (Pronzato, 2009a) that, under suitable identifiability conditions on the set $\mathcal{X}$ (supposed to be finite), $\hat\theta^n \overset{\text{a.s.}}{\to} \bar\theta$ and $\mathbf{M}_n(\hat\theta^n) \overset{\text{a.s.}}{\to} \mathbf{M}_*(\bar\theta)$, with $\mathbf{M}_*(\bar\theta)$ the D-optimal information matrix at $\bar\theta$, and one can thus take $\mathbf{C}_n = \mathbf{M}_*^{1/2}(\bar\theta)$ and c_n constant in Th. 2.

4. Sequential design and ML estimation in Bernoulli trials

4.1. Strong consistency

Consider the case of dose-response experiments with

$$Y \in \{0,1\}\,, \quad \text{with} \;\; \Pr\{Y=1|x_i,\theta\} = \eta(x_i,\theta)\,. \tag{19}$$

We suppose that Θ is a compact subset of $\mathbb{R}^p$, that $\bar{\theta}$, the 'true' value of θ that generates the observations, lies in the interior of Θ, and that $\eta(x,\theta) \in (0,1)$ for any $\theta \in \Theta$ and $x \in \mathcal{X}$. The log-likelihood for the observation Y at the design point x is given by $l(Y,x;\theta) = Y\log[\eta(x,\theta)] + (1-Y)\log[1-\eta(x,\theta)]$. We suppose that when n observations $Y_1,\ldots,Y_n$ are performed at the design points $x_1,\ldots,x_n$, the Y_i's are independent conditionally on the x_i's (so that the conditional log-likelihoods satisfy $l(Y_i|x_i,Y_{j\neq i},x_{j\neq i},\theta) = l(Y_i,x_i;\theta)$ for all i). We assume that x_i is a non-random function of $Y_1,\ldots,Y_{i-1}$, $x_1,\ldots,x_{i-1}$ for all i (as it is the case for experiments designed sequentially). The log-likelihood for n observations is then $L_n(\theta) = \sum_{i=1}^n l(Y_i,x_i;\theta)$. We denote by $\hat{\theta}^n$ the Maximum-Likelihood (ML) estimator of θ, given by $\hat{\theta}^n = \arg\max_{\theta\in\Theta} L_n(\theta)$. Although the model and estimator differ from those in Sect. 2, we obtain the following property, similar to Th. 1.

Theorem 3. *Suppose that $\mathcal{X}$ is a finite set. If $D_n(\theta,\bar{\theta})$ given by (3) satisfies (13) with $\{\tau_n\}$ a nondecreasing sequence of positive deterministic constants, then the ML estimator $\hat{\theta}^n$ in the model (19) satisfies (9).*

Proof. The first part of the proof consists in establishing that $\liminf_{n\to\infty} \inf_{\|\theta-\bar{\theta}\|\geq\delta/\tau_n}[L_n(\bar{\theta}) - L_n(\theta)] > 0$ a.s. for any $\delta > 0$ implies (9). This can be done in a way similar to the proof of Th. 1. The second part uses the following inequality (obtained by straightforward calculations)

$$L_n(\bar{\theta}) - L_n(\theta) \geq D'_n(\theta,\bar{\theta}) \left[1 - \frac{\sum_x \left|\sum_{i\in\mathcal{I}_n(x)} \zeta_i(\bar{\theta})\right| \, \{\bar{\eta}_x[1-\bar{\eta}_x]\}^{1/2} \, \left|\log\left\{\frac{\bar{\eta}_x[1-\eta_x]}{\eta_x[1-\bar{\eta}_x]}\right\}\right|}{D'_n(\theta,\bar{\theta})}\right]$$

where we denoted $\mathcal{I}_n(x) = \{i \in \{1,\ldots,n\} : x_i = x\}$, $\eta_x = \eta(x,\theta)$, $\bar{\eta}_x = \eta(x,\bar{\theta})$,

$$\zeta_i(\theta) \;\; = \;\; \frac{Y_i - \eta(x_i,\theta)}{\{\eta(x_i,\theta)[1-\eta(x_i,\theta)]\}^{1/2}}\,, \;\; i = 1,\ldots,n\,, \tag{20}$$

$$D'_n(\theta,\bar\theta) \;=\; \sum_{i=1}^{n} g[\eta(x_i,\bar\theta),\eta(x_i,\theta)]\,,$$

with $g(a,b) = a\log(a/b) + (1-a)\log[(1-a)/(1-b)]$, $(a,b)\in(0,1)^2$. (Notice that, conditionally on $x_i = x$, the random variables $\zeta_i(\bar\theta)$ are i.i.d. with zero mean and variance 1.) One can easily check that $g(a,b) > 2(a-b)^2$ with $g(a,a) = 0$, so that $D'_n(\theta,\bar\theta) \geq 2D_n(\theta,\bar\theta)$, see (3). Define $\rho(a,b) = \sqrt{a(1-a)/g(a,b)}\,\left|\log\left[\frac{a(1-b)}{b(1-a)}\right]\right|$ for $(a,b)\in(0,1)^2$. For any fixed a in $(0,1)$, $\rho(a,b)$ tends to infinity for b tending to 0 or 1 and is bounded on $(0,1)$. Straightforward calculations then give

$$\begin{aligned} L_n(\bar\theta) - L_n(\theta) \;&\geq\; D'_n(\theta,\bar\theta)\,\left[1 - \frac{1}{\sqrt{D'_n(\theta,\bar\theta)}}\sum_{x\in\mathcal{X}}\frac{\left|\sum_{i\in\mathcal{I}_n(x)}\zeta_i(\bar\theta)\right|\,\rho[\eta(x,\bar\theta),\eta(x,\theta)]}{\sqrt{r_n(x)}}\right] \\ &\geq 2D_n(\theta,\bar\theta)\,\left[1 - \frac{1}{\sqrt{2D_n(\theta,\bar\theta)}}\sum_{x\in\mathcal{X}}\frac{\left|\sum_{i\in\mathcal{I}_n(x)}\zeta_i(\bar\theta)\right|\,\bar\rho}{\sqrt{r_n(x)}}\right] \end{aligned}$$

with $\bar\rho = \sup_{x\in\mathcal{X},\,(\theta,\bar\theta)\in\Theta^2}\rho[\eta(x,\bar\theta),\eta(x,\theta)]$. Using the law of the iterated logarithm and (13) we obtain (9). ∎

4.2. Asymptotic normality

We suppose that H_η is satisfied and denote

$$\mathbf{f}_\theta(x) = \{\eta(x,\theta)[1-\eta(x,\theta)]\}^{-1/2}\frac{\partial\eta(x,\theta)}{\partial\theta}\,. \tag{21}$$

When x_i are non-random constants, the contribution of the design point x_i to the Fisher information matrix for θ is $\mu(x,\theta) = \mathbf{f}_\theta(x)\mathbf{f}_\theta^\top(x)$. Although $\mathbf{M}_n(\bar\theta)$ given by (14) is not the Fisher information matrix when the design $x_1,\ldots,x_n$ is constructed sequentially, we obtain a property similar to Th. 2 when $\mathcal{X}$ is a finite set.

Theorem 4. *Suppose that $\mathcal{X}$ is a finite set and that H_η is satisfied. If there exist non-random symmetric positive definite $p\times p$ matrices $\mathbf{C}_n$ satisfying (15), with $c_n = \lambda_{\min}(\mathbf{C}_n)$ and $D_n(\theta,\bar\theta)$ satisfying*

$$n^{1/4}c_n\to\infty \text{ and } \inf_{\|\theta-\bar\theta\|\geq c_n^2\delta} D_n(\theta,\bar\theta)/(\log\log n) \overset{\text{a.s.}}{\to} \infty \text{ for all } \delta>0 \;(n\to\infty)\,, \tag{22}$$

then the ML estimator $\hat\theta^n$ in the model (19) satisfies (17) with $\sigma^2 = 1$.

Proof. The proof is similar to that of Th. 2. It relies on a series expansion of the derivative of $L_n(\theta)$ (being now maximum at $\hat{\theta}^n$), with $\partial L_n(\theta)/\partial\theta = \sum_{i=1}^n \zeta_i(\theta)\mathbf{f}_\theta(x_i)$ and $\partial^2 L_n(\theta)/(\partial\theta\partial\theta^\top) = -n\mathbf{M}_n(\theta) + \sum_{i=1}^n \zeta_i(\bar{\theta})\left[\bar{\eta}_i(1-\bar{\eta}_i)\right]^{1/2}\mathbf{Q}_i + \sum_{i=1}^n(\bar{\eta}_i - \eta_i)\mathbf{Q}_i$, where $\mathbf{f}_\theta(x_i)$ is given by (21) and $\zeta_i(\theta)$ by (20), and where we denoted $\eta_i = \eta(x_i,\theta)$, $\bar{\eta}_i = \eta(x_i,\bar{\theta})$ and $\mathbf{Q}_i = \mathbf{Q}_i(\theta) = \left[\partial^2\eta_i/(\partial\theta\partial\theta^\top) + (2\eta_i - 1)\mathbf{f}_\theta(x_i)\mathbf{f}_\theta^\top(x_i)\right]/[\eta_i(1-\eta_i)]$. The developments are parallel to those of Th. 2, using $\|\mathbf{M}_n(\theta) - \mathbf{M}_n(\bar{\theta})\| \leq A\|\theta - \bar{\theta}\|$ for some $A > 0$, $\max_i |\bar{\eta}_i - \eta_i|\,\|\mathbf{Q}_i\| \leq B\|\theta - \bar{\theta}\|$ for some $B > 0$, $\max_i[\bar{\eta}_i(1-\bar{\eta}_i)]^{1/2}\,\|\mathbf{Q}_i\| \leq C$ for some $C > 0$ and the fact that $|\sum_{i=1,\,x_i=x}^n \zeta_i(\bar{\theta})|/\sqrt{n}$ is bounded in probability for all $x \in \mathcal{X}$. Therefore, we only require that $c_n^2\sqrt{n} \to \infty$ and $\|\hat{\theta}^n - \bar{\theta}\|/c_n^2 \xrightarrow{\text{p}} 0$ as $n \to \infty$, which follows from (22) and Th. 3. ∎

Remark 3. The condition (22) can be replaced by $\inf_{\|\theta-\bar{\theta}\|\geq c_n^2\delta} D_n(\theta,\bar{\theta}) \xrightarrow{\text{p}} \infty$ for all $\delta > 0$, $n^{1/4}c_n \to \infty$ and $\hat{\theta}^n \xrightarrow{\text{a.s.}} \bar{\theta}$. Indeed, a straightforward modification of Th. 3 shows that the first condition is enough to obtain $\|\hat{\theta}^n - \bar{\theta}\|/c_n^2 \xrightarrow{\text{p}} 0$ as $n \to \infty$.

5. Conclusions and applications

Sufficient conditions for the strong consistency and asymptotic normality of the LS estimator in nonlinear regression have been derived under the assumption that the design space is finite. Similar results apply to ML estimation in Bernoulli trials. This has important consequences for studying the asymptotic properties of nonlinear estimates in sequentially constructed experiments.

Sequential D-optimal design is considered in (Pronzato, 2009a), with the results indicated in Remark 2-(iii). Similar properties hold for adaptive *penalized* D-optimal designs for which

$$x_{n+1} = \arg\max_{x\in\mathcal{X}} \mathbf{f}_{\hat{\theta}^n}^\top(x)\mathbf{M}_n^{-1}(\hat{\theta}^n)\mathbf{f}_{\hat{\theta}^n}(x) - \gamma_n\phi(x,\hat{\theta}^n)\,, \tag{23}$$

where $\phi(x,\theta)$ denotes a penalty function related to the cost of an observation made at x. For instance, in clinical trials ϕ can be related to the probability of efficacy and no toxicity, see Dragalin and Fedorov (2006); Pronzato (2009b). A construction similar to (23) can be used for self-tuning optimization with ϕ the function of interest, to be minimized, and

$\mathbf{f}^\top_{\hat{\theta}^n}(x)\mathbf{M}_n^{-1}(\hat{\theta}^n)\mathbf{f}_{\hat{\theta}^n}(x)/\gamma_n$ playing the role of a penalty for poor estimation, see Pronzato (2000).

When γ_n in (23) is a non-random constant, under identifiability conditions on the set $\mathcal{X}$ similar to those in (Pronzato, 2009a), and assuming that $|\phi(x,\theta)|$ is bounded for all $x \in \mathcal{X}$ (finite) and $\theta \in \Theta$, we obtain that $\hat{\theta}^n$ is strongly consistent and asymptotically normal. This remains true if γ_n is a $\mathcal{F}_n$-measurable random variable (with $\mathcal{F}_n$ generated by $Y_1, \ldots, Y_n$) that tends a.s. to a non-random constant as $n \to \infty$ (in particular, one may take γ_n as a function of $\hat{\theta}_n$). Developments similar to those in (Pronzato, 2009a) show that the strong consistency of $\hat{\theta}^n$ is preserved when $\{\gamma_n\}$ is a non-random increasing sequence satisfying $\gamma_n \to \infty$ and $\gamma_n(\log\log n)/n \to 0$ in model (1) with i.i.d. errors or in model (19). For LS estimation in model (1) with $\{\varepsilon_i\}$ a martingale difference sequence, we require $\gamma_n(\log n)^\rho/n \to 0$ for some $\rho > 1$, a condition similar to that obtained in (Pronzato, 2000) when $\eta(x,\theta)$ is linear in θ (without the assumption that $\mathcal{X}$ is finite). The details will be presented elsewhere. Asymptotic normality is difficult to establish when $\gamma_n \to \infty$ since there is no obvious choice for the matrices $\mathbf{C}_n$ of Th. 2 and 4. A possible candidate is $\mathbf{C}_n = \bar{\mathbf{M}}_n^{1/2}(\bar{\theta})$ with $\bar{\mathbf{M}}_n(\bar{\theta})$ the design matrix generated by iterations similar to (23) but with $\bar{\theta}$ substituted for $\hat{\theta}^n$.

Acknowledgments

The author wishes to thank the two referees for their careful reading and comments that contributed to improve the readability of the paper.

References

Caines, P., 1975. A note on the consistency of maximum likelihood estimates for finite families of stochastic processes. Annals of Statistics 3 (2), 539–546.

Chow, Y., 1965. Local convergence of martingales and the law of large numbers. Annals of Math. Stat. 36, 552–558.

Christopeit, N., Helmes, K., 1980. Strong consistency of least squares estimators in linear regression models. Annals of Statistics 8, 778–788.

Dragalin, V., Fedorov, V., 2006. Adaptive designs for dose-finding based on efficacy-toxicity response. Journal of Statistical Planning and Inference 136, 1800–1823.

Dvoretzky, A., 1972. Asymptotic normality for sums of dependent random variables. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. II: Probability theory. Univ. California Press, Berkeley, Calif., pp. 513–535.

Hall, P., Heyde, C., 1980. Martingale Limit Theory and Its Applications. Academic Press, New York.

Jennrich, R., 1969. Asymptotic properties of nonlinear least squares estimation. Annals of Math. Stat. 40, 633–643.

Lai, T., 1994. Asymptotic properties of nonlinear least squares estimates in stochastic regression models. Annals of Statistics 22 (4), 1917–1930.

Lai, T., Robbins, H., Wei, C., 1978. Strong consistency of least squares estimates in multiple regression. Proc. Nat. Acad. Sci. USA 75 (7), 3034–3036.

Lai, T., Wei, C., 1982. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. Annals of Statistics 10 (1), 154–166.

Pronzato, L., 2000. Adaptive optimisation and D-optimum experimental design. Annals of Statistics 28 (6), 1743–1761.

Pronzato, L., 2009a. One-step ahead adaptive D-optimal design on a finite design space is asymptotically optimal. Metrika (to appear, DOI: 10.1007/s00184-008-0227-y).

Pronzato, L., 2009b. Penalized optimal designs for dose-finding. Journal of Statistical Planning and Inference (to appear, DOI: 10.1016/j.jspi.2009.07.012).

Wu, C., 1981. Asymptotic theory of nonlinear least squares estimation. Annals of Statistics 9 (3), 501–513.